DOCUMENT RESUME

ED 177 219 TM 009 746

AUTHOR Jensen, Arthur R.
TITLE Uses of Sibling Data in Psychological and Educational Research.
PUB DATE 11 Apr 79
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Age Differences; Analysis of Covariance; Cognitive Measurement; Correlation; Data Analysis; Differences; Educational Research; Environmental Influences; Family Background; *Family Influence; Genetics; Multiple Regression Analysis; Nature Nurture Controversy; *Psychological Studies; Raw Scores; *Research Design; *Research Problems; *Siblings; Standardized Tests; *Statistical Analysis; Test Bias; Tests of Significance; Trend Analysis

ABSTRACT

The uses of data on siblings for educational and psychological research purposes are explained. Examples are based on empirical data. Five types of research problems, using sibling data, are explored: (1) statistical control of family background variance in psychological or educational studies which use analysis of covariance (ANCOVA) or multiple regression analysis; (2) testing the adequacy of age-standardized scores; (3) testing the interval scale property of age standardized scores, such as intelligence or achievement quotients; (4) analysis of the correlations among variables into components which can be attributed to correlation among variables between families or to correlation among variables within families; and (5) assessment of the cultural bias in test scores or of the common factors measured by tests. Between-family-variance is defined as variance due to genetic and environmental influences common to all siblings in a family, but which differ between families. Variance within families is described as genetic and environmental influences which cause siblings to differ from one another. Some types of variables can have both between-family and within-family differences. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177219

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

USES OF SIBLING DATA IN PSYCHOLOGICAL AND EDUCATIONAL RESEARCH

Arthur R. Jensen
Institute of Human Learning
University of California, Berkeley

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Arthur Jensen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM009 746

Paper presented at the Annual Convention of the American Educational Research Association, San Francisco, April 11, 1979.

Uses of Sibling Data in Psychological and Educational Research

Arthur R. Jensen

University of California, Berkeley

ABSTRACT

Methods are explicated, with empirical examples, for using sibling data on psychometric variables as a covariate for statistically controlling family background in psychological and educational experiments, as a means for testing the adequacy of age-standardized scores, for testing the interval scale property of mental measurements, for distinguishing intrinsic from extrinsic correlations between variables, for analyzing correlations into between- and within-family components, and for detecting cultural (i.e., between families) sources of variance in psychological tests.

Uses of Sibling Data in Psychological and Educational Research

Arthur R. Jensen

University of California, Berkeley

We usually think of data obtained on related persons, family members, or various kinships in general, as pertinent only to research in genetics or behavioral genetics. Kinship data are, of course, the main grist of research in human genetics. But kinship data can also serve useful purposes in psychometrics, and in purely psychological and educational research, quite apart from any concern with genetical analysis *per se*, such as estimating the heritability of a trait or analyzing the total variance into various genetic and nongenetic components, which are among the chief purposes of quantitative genetics. We can use kinship data also to answer certain questions of strictly psychometric or psychological and educational interest.

The most plentiful and easiest kinship data to obtain in the school-age population are data on full siblings. A very substantial proportion (we have found proportions from about .50 to .70 in various California schools) of school children have one or more siblings enrolled in the same

school. Of course, the proportion of children with siblings in the same school is larger for children in the middle grades, and so that is where one should begin to recruit subjects for a sibling study.

The aim of this paper is to explicate several psychometric and psychological uses of sibling data, with examples based on data obtained by the author. The types of research problems that are addressed here by the use of sibling data fall into five general classes:

(1) Statistical control of a greater proportion of "family background" variance in psychological or educational experiments analyzed by analysis of covariance (ANCOVA) or multiple regression analysis.

(2) Testing the age-standardization of scores on tests or other measurements.

(3) Testing the interval scale property of age-standardized scores, such as IQ or AQ (achievement quotient).

(4) Analyzing the correlations among variables into components attributable to (a) correlation among variables between families and (b) correlation among variables within families -- a distinction (i.e., between a and b) which can have important theoretical implications for the interpretation of a coefficient of correlation between two (or more) variables.

(5) As a special case of #4 (above), assessing cultural bias in the test scores or the common factors measured by the items or subtests of a test or by a battery of various tests.

1. Sibling Data as a Covariate Control

The power of the test of significance of an experiment involving contrasts between independent treatment groups can be increased by means of analysis of covariance (ANCOVA). Variance in the dependent variable that is attributable to linear regression of the dependent variable on some one or more covariate control variables is removed from the within-groups (also termed "residual" or "error") variance, making for a more powerful test of the treatment effect; i.e., the experimental "error" is reduced by ANCOVA.

In educational experiments we freqeuntly wish statistically to control subject variables in order to improve the statistical detection of treatment effects. Pupils' intelligence or prior level of achievement are commonly used covariates for this purpose, as are measures of family background. Sibling data may be used as an additional measure of family background factors as they affect the particular dependent variable in the experiment, say, scores on a test of scholastic achievement. The covariance (or correlation) between the age-standardized achievement scores of the pupils in the study proper and of their siblings (preferably the siblings closest in age) represents variance due to

"family background" factors common to siblings from the same family. Or, to put it another way, it is variance due to differences between families. Using nearest-in-age siblings, measured on the same age-standardized dependent variable, as a covariate control, along with other subject variables, such as IQ or the subject's socioeconomic status, which may also be included as covariates in the ANCOVA, may increase the power of the experiment. This is more especially true in the abilities domain, where the correlations between siblings are substantial. (For example, the mean of all of the sibling correlations for IQ reported in the literature is +0.49.) Sibling data on the dependent variable is often easier to obtain than any other family background data. This can be an important consideration when there are restrictions on the invasion of privacy, which may be involved in obtaining family background data. Sibling scores on the dependent variable may serve the same purpose, controlling for family background, without the invasion of privacy implied in obtaining other family background information such as parental income, occupation, and education.

"Family background" is an ambiguous and poorly understood term in educational research. It is often mistakenly understood to mean an exclusively environmental source of variance, but in fact it is nearly always some composite of between families environmental differences and genetic differences.

Between- and Within-Family Variance. At this point, and henceforth, it is essential that the reader clearly understand the concepts of between- and within-family variance. Variance between families (ignoring measurement error) is due to those genetic and environmental influences which are common to all siblings within a family, but which differ between families. ("Family" in this context refers only to full siblings who are reared together.) Variance within families (ignoring measurement error) is due to those genetic and environmental influences which cause siblings (reared together) to differ from one another.

Some classes of variables can have both between family and within family differences. This is certainly true of all genetic factors that cause variation in the population, except in the case where the siblings are from a monozygotic multiple birth, such as identical twins or triplets, etc., for which there is no within-family genetic variance. Some types of variables, such as racial origin, have between family but no within family differences. There are still other types of variables which have no within-family variation at one period of life but have it at another period. Socioeconomic status (SES) is such an example. Siblings reared together during childhood would all have the same SES. But later, as adults, siblings may differ in SES, in terms of their differing amounts of education, occupational levels, and earnings. There are no characteristics on which there are differences within families but on which there are

no differences between families. Any genetically conditioned variable on which siblings differ will inevitably show differences between families in the next generation. This can be stated in general terms as the First Law of Differential Psychology: All within-family phenotypic and genotypic differences and correlations also exist as between-family phenotypic and genotypic differences and correlations, but the reverse is neither necessarily nor always true.

Sibling Correlation. The correlation between siblings on a given variable is entirely attributable to variance between families. The sibling correlation is, in fact, the proportion of the total variance which is attributable to variance between families. In the analysis of variance, the total variance in the population is partitioned into a between-families component σ^2_B and a within-families component σ^2_W. The population intraclass correlation between siblings, then, is $\rho_i = \sigma^2_B/(\sigma^2_B + \sigma^2_W)$. The sample intraclass correlation between siblings, derived from the between and within meansquares (MS_B and MS_W) of the analysis of variance, is $r_i = (MS_B - MS_W)/(MS_B + [\bar{k}-1]MS_W)$, where $\bar{k}$ is the arithemetic mean of the frequency distribution of the number of siblings per family. (When k, the number of siblings in each family, differs markedly among families or the frequency distribution of k is markedly skewed, it is more accurate to use either the geometric mean or the harmonic mean, instead of the arithmetic mean, for $\bar{k}$ in the above

equation.)

The sibling *intraclass* correlation r_i estimates the correlation between siblings in general, since, unlike the *interclass* or Pearsonian r, it does not assign siblings to different classes, such as x and y. The simple Pearsonian r_{xy} which would be obtained by correlating, say, younger siblings (as x) with their next older siblings (as y), or lower-scoring siblings with their higher-scoring siblings, or any other classifications of siblings as x and y, will rarely be the same as the intraclass r_i, which closely approximates the average value of all possible Pearson correlations between siblings.

In genetical research, in which we are interested in estimating the proportion of the total variance that a given kinship shares in common, the intraclass correlation is the proper measure of relationship. In certain psychometric or statistical uses of sibling (or other kinship) data, however, we should use the Pearson r. This is evident in any use of sibling data that implies a distinct classification of the members of each sibling pair, such as younger vs. older, male vs. female, higher- vs. lower-scoring, etc. Pearson r is obviously called for in the use of siblings as a covariate control in the analysis of covariance discussed above, since one member of each sibling pair is classed as the covariate in the ANCOVA. (Pearson r is, of course, implicit in the usual computational routines for ANCOVA.)

2. Testing the Age-Standardization of Measurements.

Since siblings (except twins) naturally differ in age, all of the uses of sibling data described herein require age-standardized measurements. Most published standardized tests provide age-standardized scores, which can be rigorously tested for adequacy of the age standardization by the use of sibling data, provided the sibling sample is sufficiently large (say total $N>200$) and representative of the population on which the test was standardized. If the sibling sample is not representative, this fact will be clearly revealed by the analyses described below. If the age standardization is shown by the sibling method to be inadequate and one wishes to use the sibling data for one of the other purposes explicated in this article, the measurements should be re-standardized, if possible, to remove any artifacts due to age differences between siblings. Poorly age-standardized measurements have the effect of artifactually inflating the correlation between twins (who are always the same age) and artifactually attenuating the correlation between siblings (who are always of different ages).

There are two main methods for obtaining age-standardized measurements: (a) normalized standardization, and (b) non-normalized standardization. In either method siblings are not required, but neither are they necessarily excluded.

(a) Normalized Standardization. This method is

advisable only when the total N is quite large. The total age range in the sample is divided into as many equal intervals as possible, with the limitation that no interval contain fewer than 100 subjects. The equal age intervals should not be greater than one year and need not be less than three months -- at this limit there is more advantage in having larger Ns within each age interval than in having narrower age intervals. The raw scores within each age interval are rank ordered from highest to lowest, and the ranks are then converted to percentile ranks. Using the tabled areas under the normal curve, the percentile ranks are transformed to z scores, which are normalized standardized scores. (The z scale may, of course, be subjected to any linear transformation to give the scores any mean and standard deviation one deems most convenient for one's purpose.) We have generally found that for ability and achievement tests, normalized standardized scores have the most desirable scale properties, e.g., they usually satisfy independent criteria for determining an equal interval scale. (See #3 below.)

(b) Standardized Age-regressed Scores. With this method one determines the best-fitting regression line (linear or nonlinear) of raw test scores on age (in months), using standard methods of curve fitting. A trend analysis should be done to determine if the regression of raw scores on age is significantly nonlinear.

If the regression of raw scores on age does not depart significantly from linearity, as is most often the case for ability tests in the age range from about 5 to 13, one obtains the age-regressed scores $\underline{\hat{X}}$ for each subject from the simple regression equation, i.e., $\underline{\hat{X}} = \underline{b}_{\underline{XA}}(\underline{A}-\bar{\underline{A}})+\bar{\underline{X}}$, where $\underline{b}_{\underline{XA}}$ is the coefficient of regression of raw scores on age (in months), $\underline{A}$ is the subject's age (in months), $\bar{\underline{A}}$ is the mean age of the entire sample, and $\bar{\underline{X}}$ is the mean test score in the entire sample. (The values of $\underline{\hat{X}}$ can be transformed to $\underline{z}$ scores or any other convenient transformation; or they can be converted to percentile ranks and then transformed via the tables of the normal curve, to normalized $\underline{z}$ scores or any desired linear transformation of the normalized $\underline{z}$ scores.)

If the regression of raw scores on age departs significantly from linearity, as revealed by the trend analysis described below, the age-regressed score $\underline{\hat{X}}$ for each subject can be obtained from the regression equation employing the set of statistically significant regression coefficients yielded by the trend analysis.

The trend analysis referred to above is most easily performed by means of a stepwise multiple regression analysis, in which successive powers of age (in months) are entered as the independent (predictor) variables and raw test score is the dependent variable. Powers of age (i.e.,

age, age^2, etc.) are entered stepwise into the mutiple regression analysis until the increment in R^2 is nonsignificant (at any desired level of confidence α), as determined by the usual F test, which is explicated in most statistics textbooks.

The successfulness of the age standardization of test scores can be most rigorously tested by the use of siblings. Ordinarily one can test the adequacy of the age standardization by testing the significance of the correlation r (or the multiple R, using powers of age as the predictor variables) between age and the standardized scores. The correlation should not differ significantly from zero if the age standardization is adequate. Standardized scores obtained on a large sample of siblings varying in age over the age range of the original standardization sample provide <u>two</u> independent tests of the adequacy of the age standardization. The statistical test is the Pearson correlation (or multiple R, using powers of age) between the <u>mean</u> age of sibling pairs and the test score <u>means</u> of sibling pairs (or sets of any number of siblings). This is referred to as the <u>between</u>-families correlation r_B between age and test scores. The second statistical test is the Pearson correlation (or the multiple R, using powers of age) between (a) the age difference between older (O) and younger (Y) siblings (i.e., age of O minus age of Y) and (b) the test score difference between older and younger siblings (i.e., score of O minus score of Y). This is refered to as the <u>within</u>-families

correlation r_W between age and test scores. The expected value of these correlations is zero, under the hypothesis that the test scores have been adequately age-standardized. That is to say, properly age-standardized scores should have zero correlation with age. If either r_B or r_W, or both, differ significantly from zero, the hypothesis of adequate age standardization can be rejected. The combined probabilities of r_B and r_W provide a more powerful test of the hypothesis when r_B and r_W separately have p values greater than α. (α is the size of the critical region used, or the probability of making a Type I error, i.e., rejecting the null hypothesis when it is true.) If we know the exact p values (call them p_B and p_W) of r_B and r_W, we can test the significance of the combined result using the method suggested by Fisher (1970, pp. 99-101); the value $-2(\log_e p_B + \log_e p_W)$ is distributed as chi square for 4 degrees of freedom.

A more stringent test of the hypothesis is achieved if r_B and r_W can be corrected for attenuation. The reliability of age is assumed to be perfect, and the reliability r_{XX} of the test scores is estimated in the usual way. The reliability $r_{\bar{X}}$ of the sibling pair *means*, then, is

$$r_{\bar{X}} = (r_{XX} + r_{YO})/(1+r_{YO}), \qquad (1)$$

where r_{YO} is the Pearson r between younger and older siblings. The reliability r_{O-Y} of the *differences* between older and younger siblings is

$$r_{\underline{O}-\underline{Y}} = (r_{\underline{X}\underline{X}} - r_{\underline{Y}\underline{O}})/(1-r_{\underline{Y}\underline{O}}). \qquad (2)$$

Correction for attenuation of $r_{\underline{b}}$ and $r_{\underline{W}}$, then, is $r'_{\underline{B}}=r_{\underline{B}}/r_{\underline{X}}^{1/2}$, and $r'_{\underline{W}}=r_{\underline{W}}/r_{\underline{O}-\underline{Y}}^{1/2}$.

3. Sibling Test of Interval Scale

It is often important in a psychological or educational study or experiment to have some independent evidence that the measurements constitute an equal-interval scale. For example, if we are testing the hypothesis that a particular type of instruction should produce a larger gain in achievement scores (as measured against an IQ-matched control group) for low IQ than for high IQ pupils, we cannot meaningfully interpret the lack or presence of this hypothesized interaction between IQ level and the magnitude of the experimental effect without some evidence that the dependent variable, achievement, is measured on an equal-interval scale. Usually we simply assume that the trait (in this case achievement) is normally distributed in the population, construct the test in such a way as to yield a normal distribution of scores (or normalize the scores by some suitable transformation), and then conclude that the scores constitute an interval scale.

Sibling data provide an independent test of our hypothesis that the scores that we wish to interpret as an interval scale are, in fact, an interval scale. The test, in essence, is the correlation $r_{\underline{MD}}$ between sibling means and sibling absolute differences on the measurements in

question.

Sibling means, in a random sample of the general school population, vary over a wide range on intelligence and achievement tests. With respect to our hypothesis, a sibling mean indicates the average level of the sibling pair on the measurement scale. We wish to know if this average level of the sibling pairs is significantly correlated with the absolute difference between the siblings. The expected correlation $r_{\underline{MD}}$ should be zero if the measurements (scores) are an equal-interval scale. This is a compelling inference only if there is no equally compelling theory that siblings should truly differ more in certain parts of the total range of scores than in other parts. Sibling differences are theoretically analyzable into three components of variance: (a) within-family genetic variance, (b) within-family environmental variance, and (c) error variance (i.e., unreliability of measurement). There is nothing in genetic theory that would lead to the expectation of a nonzero correlation between <u>mean</u> sibling genotypic values (which reflect between-families genetic variance) on a continuous trait and the <u>differences</u> between siblings' genotypic values (which reflect within-families genetic variance) excluding cases of major gene defects (e.g., phenylketonuria) and chromosomal anolmalies (e.g., Down's syndrome). Also, there is no <u>general</u> theoretical rationale that would lead to the expectation of a nonzero correlation between sibling means and environmental differences among siblings (i.e., within-

families environmental variance). Finally, we can empirically determine whether measurement error is homogeneous throughout the full range of the scale of scores.

The test of the hypothesis of an interval scale proceeds as follows.

(a) The scale is assumed to be properly age-standardized. The adequacy of the age-standardization can be checked by the method described in #2 above.

(b) The homogeneity of reliability can be checked, using all of the data without respect to sibling classification. The test is split into equal halves by some psychometrically sensible method, such as odd vs. even-numbered items. The Pearson r (or multiple R) between total scores and the absolute difference between the two half-scores provides a test of the homogeneity of reliability. (This is unfortunately not a powerful test and can often result in a Type II error [i.e., accepting the hypothesis $\rho=0$ when $\rho\neq0$], the risk of which can be lessened by setting the significance level for rejecting the null hypothesis at $p<.10$ or even $p<.15$.) The test is performed by the same general method as the test for interval scale described in (c) below, except that correction for attenuation is not possible, since one is actually testing the hypothesis that the reliability of the absolute difference scores is zero, i.e., that they are purely random errors which cannot be correlated with the true scores.

(c) The test for interval scale is most rigorous if one uses stepwise multiple correlation to detect any nonlinear trend, since the inequalities in score intervals may not be a simple linear function of the true scale. To simplify exposition I shall use M for the mean of a sibling pair and $|D|$ for the absolute difference between siblings. Only two siblings per family should be used, preferably those nearest in age. (Obviously, the test scores should not determine the selection of siblings.) Obtain two stepwise multiple correlations (1 and 2, below) based on the n independent variables which add significant stepwise increments to R^2.

	Independent Variables	Dependent Variable
(1)	$M, M^2, M^3 \ldots M^n$	$\|D\|$
(2)	$\|D\|, \|D\|^2, \|D\|^3 \ldots \|D\|^n$	M

The resulting multiple correlation coefficient R, in each case, should be corrected for bias, using the well-known "shrinkage" formula:

$$R_c = \sqrt{1-(1-R^2)[(N-1)/(N-n)]} \quad , \qquad (3)$$

where N is the number of sibling pairs and n is the number of independent variables. Each R_c should then be corrected for attenuation, to obtain $R_{c'} = R_c/(r_M r_{|D|})^{1/2}$, where r_M and $r_{|D|}$ are the reliabilities of the sibling means and sibling absolute differences, respectively:

$$r_M = (r_{XX} + r_i)/(1+r_i) \quad (4)$$

and

$$r_{|D|} = (r_{XX} - r_i)/(1-r_i), \quad (5)$$

where r_{XX} is the test reliability and r_i is the intraclass correlation between siblings. Finally, the two values of $R_{c'}$ are tested for significance by means of the t test:

$$t = R_{c'}N^{1/2}/(1-R^2_{c'}), \quad (6)$$

where N is the number of sibling pairs. If t does not fall in the critical region (i.e., $p>\alpha$), we conclude that the test scores are an interval scale over the range subtended by the total distribution of M.

This method was applied in a study (Jensen, 1977) of the California Test of Mental Maturity IQ scores in large sibling samples of white and black school children in rural Georgia who differed by 2σ in mean IQ. The test of interval scale was applied in each racial sample separately, and in the combined samples. In no case was the hypothesis of an equal interval scale rejected.

4. Between- and Within-Family Correlations

The total covariance between any two variables X and Y is $\Sigma xy/N$, where x and y are deviations from the mean (i.e. $x=X-\bar{X}$ and $y=Y-\bar{X}$) and N is the number of paired variables. (Pearson $r = \Sigma xy/N\sigma_x\sigma_y$.)

If n pairs of siblings, with the members of each pair

designated as a and b respectively, are each measured on variables X and Y, the total covariance between X and Y can be analyzed into two additive parts: between-families (BF) and within-families (WF) covariance. The BF covariance is the mean cross-product of the sibling means on x and y:

$$\text{BF cov} = \frac{1}{n}\Sigma\left(\frac{X_a+X_b}{2}\right)\left(\frac{Y_a+Y_b}{2}\right). \qquad (7)$$

The WF covariance is the mean cross-products of one-half of the sibling differences (i.e., the mean of the deviations of each sibling from the family mean):

$$\text{WF cov} = \frac{1}{n}\Sigma\left(\frac{X_a-X_b}{2}\right)\left(\frac{Y_a-Y_b}{2}\right). \qquad (8)$$

The total covariance between X and Y is:

$$\text{Total cov} = \frac{1}{2n}(\Sigma x_a y_a + \Sigma x_b y_b). \qquad (9)$$

By expanding equations (7) and (8) it can easily be seen that the total covariance is equal to the sum of the BF and WF covariances:

$$\text{BF cov} = \frac{1}{4n}(\Sigma x_a y_a + \Sigma x_b y_b + \Sigma x_b y_a + \Sigma x_a y_b),$$

$$\text{WF cov} = \frac{1}{4n}(\Sigma x_a y_a + \Sigma x_b y_b - \Sigma x_b y_a - \Sigma x_a y_b).$$

In working with correlations, it is simplest to use the sum of sibling scores on X and Y to obtain the BF r_{xy}, and the difference between sibling scores to obtain the WF r_{xy}. Obviously the direction of the sibling difference must be consistent for X and Y (i.e., X_a-X_b and Y_a-Y_b); it is most convenient to assign older and younger siblings to a and b,

respectively.

The BF and WF correlations should be corrected for attenuation in the usual way, using the appropriate reliabilities as given in formulas (1) and (2).

Partitioning correlatons between variables into their BF and WF correlations has theoretically important uses in differential psychology and psychometrics. It permits a separation of between-family social-cultural factors (macroenvironments) and within-family (microenvironmental) factors that contribute to the total variance in the population. The BF correlation (corrected for attenuation) reflects genetic and environmental factors that differ between families but not among siblings within families. The WF correlation (corrected for attenuation) reflects genetic and environmental factors that differ among siblings. In the absence of assortative mating (i.e., correlation between parents), the genetic component of the total variance is evenly divided between BF and WF; that is, half of the total genetic variance in the population exists between families and half of it exists within families. The increase in genetic variance due to assortative mating all goes into the BF variance; WF variance is negligibly affected by assortative mating (Jensen, 1978).

Intrinsic and Extrinsic Correlation. This theoretically important distinction for understanding correlations is revealed by BF and WF correlations.

There are four main types of what I shall refer to as intrinsic correlation:

(1) Causal-functional -- variables X and Y involve a direct causal relationship, such that experimental manipulation of X is accompanied by a change in Y. Example: number of learning trials (X) and amount of retention (Y).

(2) Common factor -- X and Y are both measures of (or are both correlated with) some common factor. Example: strength of left hand grip (X) and strength of right hand grip (Y); or height (X) and weight (Y); or speed of learning X and speed of learning Y.

(3) Part-whole -- one variable is some part of the other, or skill X is a subset of skill Y. Example: leg length (X) and height (Y); or skill in short division (X) and skill in long division (Y).

(4) Pleiotropy -- X and Y are both affected by the same gene(s), even though X and Y may be phenotypically quite distinct characteristics. Variation in both traits is linked to a common (genetic) causal factor. A pleiotropic gene has two (or more) distinct phenotypic effects. Example: phenylketonuria, a single-locus genetic defect, results in mental retardation (X) and lightness of skin and hair pigmentation (Y).

There are two types of extrinsic correlation.

(1) Nonlinked genetic correlation -- X and Y can be correlated in the population through common assortment of genes due to cross-assortative mating for certain characteristics which have no functional or other intrinsic relationship to one another. Say, for example, there is zero correlation between curly hair and height in the population, but in the next generation there is a strong tendency for tall men to marry curly-haired women (i.e., cross-assortative mating for the two characteristics); then, in the next generation there will be a genetic correlation between height and curly hair in the population.

Such cross-assortative mating can create a correlation in the population between any genetically conditional traits, physical or psychological, even though there may be no intrinsic relationship between them. Such genetic correlations (unless there is genetic linkage, which is highly improbable for continuous or polygenic traits) have the property that they are entirely between-families (BF) correlations and have no within-families (WF) component. Even though there is reliable WF variance on each trait, the expected value of the WF correlation between the traits is zero. A test of the hypothesis that WF $\rho_{xy}=0$, when the BF $\rho>0$, decisively rules out genetic linkage between traits X and Y as well as any form of intrinsic correlation between them. If WF $\rho_{xy}=0$, the total population $\rho_{xy}>0$ (i.e., for all individuals in the population) represents entirely extrinsic correlation, due either to (a) genetic correlation

(resulting from cross-assortative mating), or to (b) the extrinsic component of what I refer to as environmental correlation (see below), or to a combination of (a) and (b).

The fact that a correlation is said to be extrinsic according to this criterion does not make it any less real or reduce its predictive validity for individuals in the general populations. The distinction between intrinsic and extrinsic correlation, however, does imply different theoretical interpretations of the correlations so classified. As I have explained elsewhere (Jensen, 1979), experimental or so-called process analysis of correlated tests, or of test scores and certain physical measurements, can lead to a greater understanding of the correlated variables only if there is an intrinsic correlation between them. For example, the correlation between height and IQ appears to be entirely extrinsic. I and others (e.g., Laycock & Taylor, 1965) have found a significant positive BF correlation (ranging from about +0.1 to +0.3) between height and IQ, while the WF correlation is zero. The fact that the correlation between height and IQ is an extrinsic correlation, as indicated by these findings, implies that individual study of the nature of variation in height can afford no scientific clues to the nature of individual variation in IQ. The discovery of physical correlates of mental traits can possibly lead to a greater understanding of the mental traits, provided the correlation is intrinsic, i.e., a WF correlation which is not due merely to genetic linkage. We

know, for example, that there is a correlation of about +0.30 between IQ and brain size (VanValen, 1974), but we do not yet know if this is an intrinsic correlation. It would seem important to find out. If it is only an extrinsic correlation we can forget it, as far as its playing any part in a theory of intelligence is concerned. The same thing can probably be said about the correlations between IQ and a number of of different blood types (Osborne & Suddick, 1971). The positive correlation between lightness of skin pigmentation and IQ in the American black population (studies reviewed by Jensen, 1973, pp. 222-224) may or may not be an intrinsic correlation; no one has yet determined the WF and BF correlations between IQ and skin color. If the WF correlation is zero, it would rule out the hypothesis which explains the observed correlation in the black population in terms of adverse effect on IQ of social prejudice against darker skin. The counter hypothesis would be that the correlation is entirely BF due to cross-assortative mating for skin color and IQ within the black population -- a finding which might be of sociological interest but not of any importance in terms of psychological or genetical theory.

Another example is the negative correlation between IQ and delinquency. This turns out to be an intrinsic or WF correlation. Delinquent and nondelinquent siblings of the same sex differ almost as much in IQ as do unrelated delinquents and nondelinquents in the general population (Hirschi & Hindelang, 1977). This implies that IQ somehow mediates

delinquency -- a quite different (and much more important) theoretical implication than would be the case if only a BF correlation (but not a WF correlation) were found between delinquency and IQ.

(2) Correlation by genetic linkage -- variables X and Y may be correlated WF only because they are genetically linked, i.e., the genes influencing X and Y are located on the same chromosomes in close proximity to one another, so that they have a higher probability of remaining together in the crossing-over process in gametogenesis (the meiotic formation of the sex cells). The two traits, however, may not be any more intrinsically correlated than in the case of nonlinked genetic correlation; and as the linkages break up in successive generations, the WF correlation (and the BF correlation, too) will gradually decrease toward zero. Linkage would account for little, if any, of the correlation between continuous or polygenic traits, but would have to be considered, for example, in the case of correlations between particular single-gene blood groups and IQ. The finding of such a genetic linkage would not be psychologically informative about the nature of IQ, but it would be of considerable importance in the study of the genetic inheritance of intelligence. There are methods in quantitative genetics for the statistical detection of linkage (e.g., Cavalli-Sforza & Bodmer, 1971, pp. 870-880).

Environmental correlation: Intrinsic and Extrinsic.

This is a class of correlations which may be either intrinsic or extrinsic; these two aspects show up in the WF and BF components, respectively, of the overall correlation in the population. If classes of events X and Y occur together more than chance in the individual's environment, through common interest or exposure, then knowledge of X probabilistically (but not causally) implies knowledge of Y. Example: A person who knows a lot about, say, baseball is also likely to possess above-average knowledge of football, through common interest in sports. An opera lover is likely to have more than average knowledge about symphonies.

Such environmental correlations can exist WF as well as BF. Sex differences in tests of various kinds of information are an example. There is a WF correlation between knowledge about sewing (x) and cooking (Y), and between knowledge about sports (X) and auto mechanics (Y), largely because of the different experiences of males and females. One would expect these WF correlations to be higher in families with opposite-sex siblings than in families with same-sex siblings.

Certain kinds of environmental experiences may be much more highly associated for all siblings of some families than for all siblings of other families, because of differing family interests, life styles, cultures, etc. Measurements of the knowledge or skills derived from much associated experiences X and Y will then show a much higher BF

correlation than a WF correlation. Marked differences between BF and WF correlations, and particularly a difference between the pattern of intercorrelations or the factor structure of the BF and WF intercorrelations of a number of variables, indicates that at least some of the intercorrelations among the variables are either extrinsic environmental correlations, or extrinsic genetic correlations, or both. Different social classes or racial groups, for example, might have different commonly associated experiences, which generally affect all members of a family as well as most families in the particular group. The WF and BF correlations between measurements that reflect these experiences will therefore be different in different subpopulations. If variance in mental test scores were largely the result of differences in social class, cultural background, economic privilege, parental education, family values, and the like, as is so often claimed by the critics of mental tests, then we should expect most of the significant intercorrelations among such tests to be BF rather than WF, and it would seem reasonable to expect different patterns (or factor structures) of BF and WF intercorrelations among various tests, and in different subpopulations.

5. Sibling Test of Culture Bias

The hypothesis that a test is not culture biased for two or more subpopulations can be rejected if it can be demonstrated, at some acceptable level of statistical signi-

ficance, that any essential psychometric characteristic of the test differs between subpopulations. The one characteristic that is excluded, of course, is the mean, since tests are intended to discriminate among individuals, and as different subpopulations are comprised of individuals, there is no reason, in principle, why the mean test scores of the subpopulations should not also differ reliably if the subpopulations are not assumed to be perfectly random samples of the total population. Every failure to reject the null hypothesis (i.e., that the subpopulations do not differ significantly on some essential psychometric characteristic of the test, such as reliability, validity, item characteristic curves, factor structures, etc.) strengthens the presumption that the test is not culturally biased for the subpopulations in question. Sibling data on a battery of tests provide an additional means for testing this null hypothesis.

This use of sibling data is illustrated by measurements obtained on pairs of siblings from 1,495 white families and 901 black families in grades 2 to 6 (ages of about 7 to 12 years) in California schools. In all cases, the pair of siblings in each family nearest in age and enrolled in grades 2 to 6 were selected for this study. In additon to measurements of height and weight, scores on the following tests were obtained:

Memory: a composite score based on three highly intercorrelated tests of rote memory involving immediate recall,

delayed recall, and learning through repetition.

Figure Copying; copying ten geometric forms of increasing difficulty (complexity).

Pictorial IQ: Lorge-Thorndike Primary Level IQ test, a nonreading test employing pictorial materials and orally group-administered.

Nonverbal IQ: Lorge-Thorndike Nonverbal, a completely nonreading, nonverbal test employing pictorial and abstract figural material.

Verbal IQ: Lorge-Thronkdike Verbal, a verbal test consisting of word similarities, opposites, verbal analogies, verbal reasoning, etc.

Vocabulary: Word Meaning subtest of the Stanford Achievement Test.

Reading Comprehension: Paragraph Meaning subtest of the Stanford Achievement Test.

Age Standardization. Raw scores on each of the above tests, as well as the height and weight measurements, were age-standardized (i.e., normalized standard scores within 6 month intervals) on the data of an entire school district with approximately 8000 pupils, 60% whites and 40% blacks. Age standardization was done separately for blacks and whites. The adequacy of the age-standardization was tested by the sibling method described under #2 above. The

shrunken multiple R between powers of sibling age differences and sibling standardized score differences was not significantly greater than zero for any of the measurements for either whites or blacks. In other words, age variance in scores was effectively removed by the standardization procedure.

Interval Scale. The age-standardized scores on each of the tests was checked for interval scale within each racial group by the method described in #3 above. All of the tests met the criterion of interval scale for both racial groups.

Height and Weight. The BF and WF correlations of height and weight (and corrected for attenuation in parentheses), and the sibling correlations, $r_{HH'}$ and $r_{WW'}$, are as follows:

		White	Black
Between Family	r_{HW}	.73(.75)	.79(.81)
Within Family	r_{HW}	.68(.71)	.70(.74)
Between Siblings	$r_{HH'}$	.42(.43)	.45(.46)
Between Siblings	$r_{WW'}$	.38(.39)	.37(.38)

Height and weight are obviously intrinsically correlated -- an example of correlation through a common factor, viz., general body size.

Height and weight show only two significant ($p<.05$) WF correlations with the seven mental tests for both whites and blacks, while there are twelve significant BF correlations

with the physical measurements, suggesting that the physical measurements are not intrinsically correlated with mental ability. The average WF and BF correlations between the physical and mental measurements are +.02 (n.s.) and +.10 ($p<.01$), respectively.

Factor Analysis of BF and WF Correlation Matrices. The BF and WF matrices of corrected correlations among the seven tests were factor analyzed separately, for whites and blacks. The general factor common to all of the tests is represented by the first principal component, which was extracted from the BF and WF correlation matrices for whites and blacks.

If the tests' intercorrelations are intrinsic rather than extrinsic, the same general factor g should appear in both the BF and WF matrices, which would indicate that test score differences between siblings reflect the same general factor as test score differences between children from different families. And if the tests are not culturally biased, one should expect to find the same g factor in whites and blacks, for both BF and WF correlations.

Table 1 shows the loadings on the first principal

Insert Table 1 about here
---

component extracted from the four correlation matrices. It can be seen that the four factors are highly similar. The

size of the loadings are generally larger for the BF than the WF first principal component, which should be expected if there is assortative mating for g, since all of the population variance attributable to assortative mating is BF variance. Also, there is undoubtedly some BF environmental variance, which appears to be similar in nature to WF environmental variance, considering the similarity between the BF and WF g factors shown in Table 1.

A quantitative index, ranging from -1 to +1, commonly used to measure the degree of similarity between factors extracted from the same set of variables in different samples is the congruence coefficient r_c (Cattell, 1978, pp. 251-255):

$$r_c = \Sigma b_i b_j \; / \; (\Sigma b_i^2 \Sigma b_j^2)^{1/2}, \qquad (10)$$

where b_i and b_j are the factor loadings on the same tests in groups i and j. The congruence coefficients among the four g factors are shown in Table 2. They are all very high and

---
Insert Table 2 about here
---

do not differ significantly, indicating that this battery of tests measures the same g factor both within and between families, and in both whites and blacks in this California school population.

Table 1

Loadings on the First Principal Component (Corrected for Attenuation) of the Between- and Within-Family Correlations for Whites and Blacks

Test	White		Black	
	Between	Within	Between	Within
Memory	.585	.285	.477	.398
Figure Copying	.492	.350	.645	.316
Pictorial IQ	.950	.789	.847	.945
Nonverbal IQ	.904	.892	.932	.871
Verbal IQ	.952	.981	.958	.924
Vocabulary	.825	.713	.698	.657
Reading Comprehension	.878	.737	.796	.688
Percent of Variance	66.6	51.9	61.0	52.4

Table 2

Congruence Coefficients Between First Principal Components of Whites (W) and Blacks (B), Between (b) and Within (w) Families

Principal Component	2) W w-f	3) B b-f	4) B w-f
1) White between-families	.987	.993	.991
2) White within-families		.986	.993
3) Black between-families			.985
4) Black within-families			

References

Cattell, R.B. The Scientific Use of Factor Analysis in Behavioral and Life Sciences. New York: Plenum, 1978.

Cavalli-Sforza, L.L., & Bodmer, W.F. The Genetics of Human Populations. San Francisco: W.H. Freeman, 1971.

Fisher, R.A. Statistical Methods for Research Workers. 14th ed. New York: Hafner (Macmillan), 1970.

Hirschi, T., & Hindelang, M.J. Intelligence and delinquency: A revisionist review. American Sociological Review, 1977, 42, 571-587.

Jensen, A.R. Cumlative deficit in IQ of blacks in the rural South. Developmental Psychology, 1977, 13, 184-191.

Jensen, A.R. Genetic and behavioral effects of nonrandom mating. In R.T. Osborne, C.E. Noble, & N. Weyl (Eds.) Human Variation: Psychology of Age, Race, and Sex. New York: Academic Press, 1978.

Jensen, A.R. g: Outmodeled theory or unconquered frontier? Creative Science and Technology, 1979, 2, 16-29.

Laycock, F., & Caylor, J.S. Physiques of gifted children and their less gifted siblings. Child Development, 1964, 35, 63-74.

Osborne, R.T., & Suddick, D.E. Blood type gene frequency and mental ability. Psychological Reports, 1971, 29, 1243-1249.

VanValen, L. Brain size and intelligence in man. American Journal of Physical Anthropology, 1974, 40, 417-423.